Computer Science Department

TECHNICAL REPORT

THREE METHODS FOR REFINING ESTIMATES OF INVARIANT SUBSPACES

By

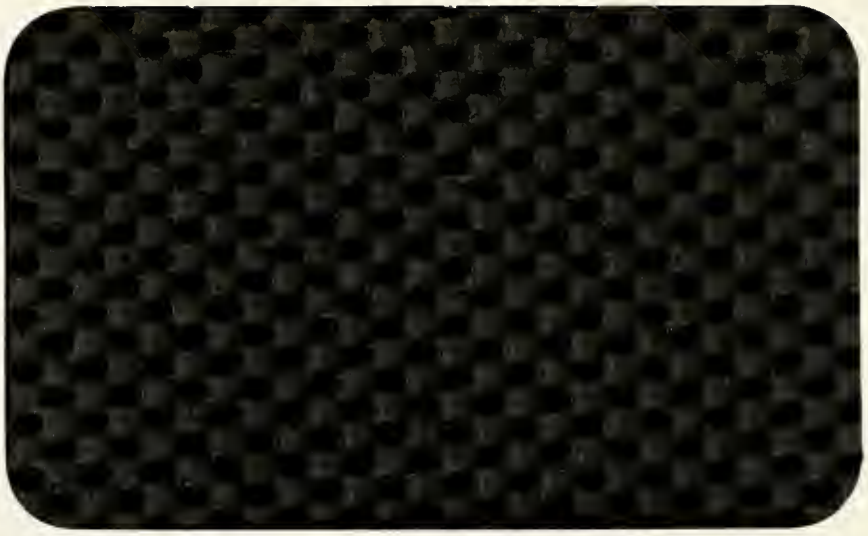
JAMES DEMMEL

OCTOBER 1985

TECHNICAL REPORT #185

NEW YORK UNIVERSITY

Department of Computer Science
Courant Institute of Mathematical Sciences
251 MERCER STREET, NEW YORK, N.Y. 10012



THREE METHODS FOR REFINING ESTIMATES OF INVARIANT SUBSPACES

By

JAMES DEMMEL

OCTOBER 1985

TECHNICAL REPORT #185

Three Methods for Refining Estimates of Invariant Subspaces

James Weldon Demmel
Courant Institute of Mathematical Sciences
251 Mercer Str.
New York, NY 10012

Abstract

We compare three methods for refining estimates of invariant subspaces, due to Chatelin, Dongarra/Moler/Wilkinson, and Stewart. Even though these methods all apparently solve different equations, we show by changing variables that they all solve the same equation, the Riccati equation. The benefit of this point of view is threefold. First, the same convergence theory applies to all three methods, yielding a single criterion under which the last two methods converge linearly, and a slightly stronger criterion under which the first algorithm converges quadratically. Second, it suggest a hybrid algorithm combining advantages of all three. Third, it leads to algorithms (and convergence criteria) for the generalized eigenvalue problem. These techniques are compared to techniques used in the control systems community.

1. Introduction.

Methods for refining estimates of invariant subspaces of matrices have been suggested in [Chatelin], [Dongarra,Moler,Wilkinson], and [Stewart]. These three methods (henceforth called C, DMW, and S, respectively), all solve apparently different equations, since they represent the desired invariant subspace slightly differently. However, by a simple change of basis, we will see that all three methods are attempting to solve the same equation, the Riccati equation:

$$AR - RB = C + RDR$$

for R, which represents the error in the initial estimate of the invariant subspace.

The benefit of this unified point of view is threefold. First, it allows the same convergence theory to be applied to all three methods, yielding the same criterion for linear convergence of DMW and S, and a slightly stronger criterion for quadratic convergence of C. Second, it suggests a hybrid algorithm (Algorithm 3 below) combining advantages of DWM, C and S. Third, it suggests analogous algorithms (and convergence theory) for the generalized eigenproblem.

The Riccati equation (and its variations) have been a central object of study in the control systems community for some time, and a number of algorithms have been proposed and implemented [Arnold,Laub]. The algorithms used there mirror the ones in the numerical analysis community: converting the Riccati equation to an eigenvalue problem and using the QR or QZ algorithms, followed by Newton iteration to refine the solution if necessary.

This paper is organized as follows. After introducing some notation in section 2, we show how the methods DWM, C and S all reduce to solving the Riccati equation in section 3. Section 4 discusses a linear iteration and Newton iteration for solving the Riccati equation, and gives convergence criteria. Section 5 presents three generalizations of DWM, C and S for solving the generalized eigenproblem, and shows how they may be again reduced to solving a generalized Riccati equation. Section 6 generalizes the results of section 4 to solving this generalized equation. Section 7 discusses the relative costs of the linear iteration and Newton, and suggests a hybrid algorithm which may be seen as a form of modified Newton. Section 8 discusses future work: a numerical comparison of DWM, C and the hybrid algorithm presented below.

2. Notation

The n by n matrix whose invariant subspace we desire will be called A. The invariant subspace we seek will be m dimensional. We will measure errors using the 2-norm for matrices

$$||T|| = \sup_{x \neq 0} ||Tx|| / ||x||$$

($||x||$ denotes the 2-norm for vectors) and the Frobenius norm

$$||T||_F = \left(\sum_{ij} |T_{ij}|^2\right)^{1/2} .$$

3. Reduction to the Riccati Equation

We first consider the method C of Chatelin. She seeks an n by m matrix X whose columns span the desired invariant subspace. Her system of equations depends on a fixed full rank n by m matrix Z:

$$AX - X(Z^*AX) = 0 \quad , \quad Z^*X = I \qquad \text{(C)}$$

Now change basis so that (in the new basis) $Z=[I|0]^T$. If Z consists of orthonormal columns

(and there is no reason it should not), then this change of basis can even be orthonormal. We will also call the new transformed matrix A, and write it as the partitioned matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} is m sub m and A_{22} is $n-m$ by $n-m$. In this coordinate system it is easy to see $Z^*X=I$ implies X is of the form

$$X = \begin{bmatrix} I \\ R \end{bmatrix}$$

where R is an arbitrary $n-m$ by m matrix. This lets us rewrite (C) as

$$A_{21} + A_{22}R - R(A_{11} + A_{12}R) = 0$$

or

$$A_{22}R - RA_{11} = -A_{21} + RA_{12}R \qquad \text{(C')}$$

the promised Riccati equation.

In the DMW method Dongarra, Moler and Wilkinson try to solve the equation

$$AX = XB \qquad \text{(DWM)}$$

simultaneously for the n by m matrix X and the m by m matrix B. Since this is nm equations in $nm+m^2$ unknowns, X is constrained by holding m of its rows fixed, reducing the number of unknowns to nm. The m rows of X to hold fixed are chosen to be the "most nonsingular" m by m submatrix of the initial approximation to X. In practice, these may be determined by LU with pivoting. If the initial approximation of X has orthonormal columns, by an orthonormal change of basis we may assume that it consists of the first m columns of the identity matrix. Thus in this new basis X may be written

$$X = \begin{bmatrix} I \\ R \end{bmatrix}$$

with R arbitrary as before, allowing us to rewrite (DWM) as

$$\begin{bmatrix} A_{11} + A_{12}R \\ A_{21} + A_{22}R \end{bmatrix} = \begin{bmatrix} B \\ RB \end{bmatrix} .$$

Substituting the first equation for B into the second yields

$$A_{22}R - RA_{11} = -A_{21} + RA_{12}R \ , \qquad \text{(DWM')}$$

the same Riccati equation as before.

Finally we consider Stewart's method S. Actually, S was originally presented not as an algorithm but as a technique for doing perturbation theory for invariant subspaces. Nonetheless, it works as an algorithm as well, and the perturbation theory derived by Stewart will later be used to derive convergence criteria for all three algorithms. S begins with an n by m matrix X spanning an approximate invariant subspace and a n by $n-m$ matrix X' spanning a complementary subspace, so that the matrix $[X|X']$ is nonsingular. The true invariant subspace is represented as $X + X'R$, where R is $n-m$ by m as before, and an equation is derived for R as follows. $X+X'R$ will be invariant if and only if the lower left $n-m$ by m block of

$$[X+X'R|X']^{-1} \cdot A \cdot [X+X'R|X'] = \left([X|X'] \cdot \begin{bmatrix} I & 0 \\ R & I \end{bmatrix} \right)^{-1} \cdot A \cdot \left([X|X'] \cdot \begin{bmatrix} I & 0 \\ R & I \end{bmatrix} \right)$$

is zero. As before, we change basis so that $[X|X']$ is the identity matrix. If X consists of

orthonormal columns and X' spans the orthogonal complement of the space spanned by X, this change of basis can be orthonormal. In this new basis we get that the lower left $n-m$ by m corner of

$$\begin{bmatrix} I & 0 \\ R & I \end{bmatrix}^{-1} \cdot \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \cdot \begin{bmatrix} I & 0 \\ R & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ -R & I \end{bmatrix} \cdot \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \cdot \begin{bmatrix} I & 0 \\ R & I \end{bmatrix}$$

must be zero, or

$$A_{22}R - RA_{11} = -A_{21} + RA_{12}R \quad , \tag{S}$$

the same Riccati equation as before.

4. Solving the Riccati Equation

There are two basic approaches to solving the Riccati equation, the iteration

$$A_{22}R_{l+1} - R_{l+1}A_{11} = -A_{21} + R_lA_{12}R_l \quad , \quad R_0=0 \tag{Iter}$$

and Newton's method. DWM and S use iteration, and C uses Newton. It turns out that Stewart's perturbation analysis yields a useful convergence criterion for (Iter) in terms of simply computable properties of the blocks A_{ij}. Stewart's analysis appears to be repeated in part in the later paper by Dongarra, Moler, Wilkinson. (Since Dongarra, Moler and Wilkinson actually solve (DWM) by a Gauss-Seidel type of iteration where newly updated components of X and B are used to update later components, their analysis is slightly different.) A very similar analysis yields almost the same criterion for (C) to have quadratic convergence. This quadratic convergence criterion does not appear in Chatelin's paper and seems to be stronger than her results.

To state the results we need to define the separation of two matrices $\mathrm{sep}(A_{11},A_{22})$ [Stewart]: it is the smallest singular value of the linear operator which maps R to $A_{22}R - RA_{11}$. It is nonzero (i.e. the operator is invertible) as long as A_{11} and A_{22} have disjoint spectra, and also satisfies $\mathrm{sep}(A,B) = \mathrm{sep}(B,A)$. One can also show that it is the smallest singular value of the $m(n-m)$ by $m(n-m)$ matrix

$$I_m \otimes A_{22} - A_{11}^T \otimes I_{n-m} \quad .$$

Theorem 1: [Stewart] Let

$$\kappa = \frac{\|A_{12}\|_F \cdot \|A_{21}\|_F}{\mathrm{sep}^2(A_{11},A_{22})} \quad .$$

Then if

$$\kappa < \frac{1}{4} \tag{LinearConvergence}$$

(Iter) converges linearly to a solution R which is the unique solution inside the ball

$$\|R\|_F \le \frac{1-(1-4\kappa)^{1/2}}{2\kappa} \cdot \frac{\|A_{21}\|_F}{\mathrm{sep}(A_{11},A_{22})} < 2 \cdot \frac{\|A_{21}\|_F}{\mathrm{sep}(A_{11},A_{22})} \quad .$$

Furthermore, the convergence is linear with a contraction constant of no more than

$$1-(1-4\kappa)^{1/2} \quad .$$

The parameter κ can be interpreted as follows. Its numerator, $\|A_{12}\|_F \cdot \|A_{21}\|_F$, measures the quality of the initial approximate invariant subspace: it will be small when the approximation is good, and the factor $\|A_{21}\|_F$ will be zero if and only if the initial approximation is in fact correct. $\|A_{12}\|_F$ will be small if X' (in S) is also a good approximate invariant

subspace. The denominator sep measures the separation of the spectra of A_{11} and A_{22}. If sep is small it means some eigenvalues of A_{11} and A_{22} can be made to merge with small changes in both A_{ii}; this means the invariant subspaces belonging to the two parts of the spectrum are unstable and hard to compute. Thus, κ will be small if we start with a good initial approximate invariant subspace and if the eigenvalues associated with that subspace are well separated from the remainder of the spectrum.

Now we turn to the quadratic convergence of C. As in [Chatelin], it is easy to see that Jacobean of $F(X) = AX - X(Z^*AX)$ with respect to X is the linear map which maps S to $DF(X)(S) = (I-XZ^*)AS - S(Z^*AX)$. The standard Newton iteration would then be

$$X_{k+1} = X_k - DF(X_k)^{-1}F(X_k)$$

At first glance this seems problematic since there are no guarantees that, first, the condition $Z^*X_k=I$ is fulfilled for all iterations k, and, second, that $DF(X_k)$ is nonsingular for all k. In fact, if X is the true invariant subspace, then $DF(X)$ will be singular if Z^*AX is singular, and if the spectrum whose corresponding subspace we seek contains zero, this will be the case. Chatelin deals with these two problems by first showing that if $Z^*X_0=I$, then all subsequent iterates also satisfy $Z^*X_k=I$, and second assuming that by shifting A to $A+\sigma I$ singularity of Z^*AX can be avoided.

We show that the same change of basis we used above leads to a simplified Newton iteration which not only guarantees $Z^*X_k=I$ but eliminates the artificial restriction on the spectrum of Z^*AX. This will also lead to a new convergence criterion only slightly more restrictive than (LinearConvergence) above. Let $Z = [I|0]$ as above and write

$$X_k = \begin{bmatrix} I \\ R_k \end{bmatrix} .$$

From this it follows straightforwardly that $Z^*F(X_k)=0$. Also in this coordinate system

$$DF(X_k)\left(\begin{bmatrix} S_1 \\ S_2 \end{bmatrix}\right) = \begin{bmatrix} -S_1(A_{11}+A_{12}R_k) \\ (-R_kA_{11}+A_{21})S_1 + (-R_kA_{12}+A_{22})S_2 - S_2(A_{11}+A_{12}R_k) \end{bmatrix}$$

from which we see that in solving

$$DF(X_k)\left(\begin{bmatrix} S_1 \\ S_2 \end{bmatrix}\right) = F(X_k) = \begin{bmatrix} 0 \\ A_{22}R_k - R_kA_{11} + A_{21} - R_kA_{12}R_k \end{bmatrix}$$

we may take $S_1=0$. This corresponds to Chatelin's statement that $DF(X)^{-1}$ maps the space of matrices M such that $Y^*M=0$ into itself. Since $R_{k+1}=R_k-Z_2$, a little more manipulation leads to the iteration

$$(A_{22}-R_kA_{12})R_{k+1} - R_{k+1}(A_{11}+A_{12}R_k) = -A_{21} - R_kA_{12}R_k \ , \qquad \text{(Newton)}$$

a set of linear equations for R_{l+1}. This is the same iteration one gets applying Newton to $A_{22}R - RA_{11} + A_{21} - RA_{12}R = 0$ directly.

The next theorem states a condition under which (Newton) converges:

Theorem 2: Let κ be defined as in Theorem 1. Then if

$$\kappa < \frac{1}{12} \ , \qquad \text{(QuadraticConvergence)}$$

(Newton) will converge quadratically to the unique solution R inside the ball given in Theorem 1. If $E_i = R_i - R$ is the error at the i-th step of the iteration, then

$$||E_{k+1}||_F \leq \frac{||A_{12}||_F}{\text{sep}(A_{11},A_{22})} \cdot \frac{3}{2} \cdot ||E_k||_F^2 \ .$$

Proof: Take (Newton), subtract from it the same equation with the solution R substituted for R_k and R_{k+1}, and rearrange to obtain

$$(A_{22}-R_kA_{12})E_{k+1} - E_{k+1}(A_{11}+A_{12}R_k) = -E_kA_{12}E_k$$

implying

$$||E_{k+1}||_F \leq \frac{||E_k||_F^2 \cdot ||A_{12}||_F}{\text{sep}(A_{11}+A_{12}R_k,A_{22}-R_kA_{12})} \qquad (*)$$

Thus quadratic convergence is evident, as long as the denominator is bounded away from 0. To prove this we need bounds on the growth of $||R_k||_F$.

From (Newton) we see that

$$||R_{k+1}|| \leq \frac{||A_{21}||_F + ||A_{12}||_F \cdot ||R_k||^2}{\text{sep}(A_{11}+A_{12}R_k,A_{22}-R_kA_{12})}$$

From [Stewart] we have the lemma that

$$\text{sep}(A+E,B+F) \geq \text{sep}(A,B) - ||E||_F - ||F||_F \qquad \text{(Lemma)}$$

which implies that

$$||R_{k+1}|| \leq \frac{||A_{21}||_F + ||A_{12}||_F \cdot ||R_k||^2}{\text{sep}(A_{11},A_{22}) - 2\,||A_{12}||_F\,||R_k||_F} \ .$$

Abbreviating

$$||R_k||_F = \frac{f_k \cdot ||A_{21}||_F}{\text{sep}(A_{11},A_{22})}$$

we see that

$$f_{k+1} \leq \frac{1 + \kappa f_k^2}{1 - 2\kappa f_k}$$

with $f_0=0$. It is easy to see that this recurrence for f_k increases monotonically and approaches $(1-(1-12\kappa)^{1/2}) \,/\, 6\kappa$ as a limit. Thus

$$||R_k||_F < \frac{1 - (1-12\kappa)^{1/2}}{6\kappa} \cdot \frac{||A_{21}||_F}{\text{sep}(A_{11},A_{22})} \ .$$

Substituting this into (*) above and using (Lemma) we have

$$||E_{k+1}||_F \leq ||E_k||_F^2 \cdot \frac{3\,||A_{12}||_F}{\text{sep}(A_{11},A_{22}) \cdot (2 + (1-12\kappa)^{1/2})}$$

which concludes the proof. Q.E.D.

Thus we see that Stewart's framework provides a convenient tool for analyzing convergence. We also see that the two convergence criteria (LinearConvergence) and (QuadraticConvergence) are almost the same, (QuadraticConvergence) being slightly more restrictive.

It is possible to make a matrix interpretation of Newton's method. It is equivalent to the following algorithm:

Algorithm 1:

1) Given bases X and X' for an approximate invariant subspace and a complementary space, transform the problem so that $[X|X']=I$.

2) Take one step of (Iter) with $R_0=0$ yielding R, replace X by the better estimate $X+X'R$ and return to step 1).

In matrix notation this algorithm produces a sequence $R^{(k)}$ which determine a sequence of similarity transforms such that

$$\begin{bmatrix} I & 0 \\ -R^{(k)} & I \end{bmatrix} \cdots \begin{bmatrix} I & 0 \\ -R^{(1)} & I \end{bmatrix} A \begin{bmatrix} I & 0 \\ R^{(1)} & I \end{bmatrix} \cdots \begin{bmatrix} I & 0 \\ R^{(k)} & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ -\sum_{l=1}^{k} R^{(l)} & I \end{bmatrix} A \begin{bmatrix} I & 0 \\ \sum_{l=1}^{k} R^{(l)} & I \end{bmatrix}$$

converges to a matrix with the lower left $n-m$ by m corner equal to 0. $\sum_{l=1}^{k} R^{(l)}$ is the same as the R_k generated by Newton.

We may see this as follows. Obviously $R^{(1)}$ is the same as the R_1 produced by (Newton). Now assume by induction that $\sum_{l=1}^{k} R^{(l)} = R_k$. Abbreviate $\sum_{l=1}^{k} R^{(l)}$ by R^{Σ}. Executing step 1) amounts by induction to transforming A to

$$\begin{bmatrix} I & 0 \\ -R^{\Sigma} & I \end{bmatrix} A \begin{bmatrix} I & 0 \\ R^{\Sigma} & I \end{bmatrix} = \begin{bmatrix} A_{11}+A_{12}R^{\Sigma} & A_{12} \\ -R^{\Sigma}A_{11}-R^{\Sigma}A_{12}R^{\Sigma}+A_{21}+A_{22}R^{\Sigma} & A_{22}-R^{\Sigma}A_{12} \end{bmatrix}.$$

Executing step 2) means solving

$$(A_{22}-R^{\Sigma}A_{12})R^{(k+1)} - R^{(k+1)}(A_{11}+A_{12}R^{\Sigma}) = R^{\Sigma}A_{11}+R^{\Sigma}A_{12}R^{\Sigma}-A_{21}-A_{22}R^{\Sigma}$$

or

$$(A_{22}-R^{\Sigma}A_{12})(R^{\Sigma}+R^{(k+1)}) - (R^{\Sigma}+R^{(k+1)})(A_{11}+A_{12}R^{\Sigma}) = -A_{21}-R^{\Sigma}A_{12}R^{\Sigma}$$

for $R^{(k+1)}$. Comparing with (Newton) shows $R^{(k+1)}+R^{\Sigma} = \sum_{l=1}^{k+1} R^{(l)} = R_{k+1}$ as desired.

5. The Generalized Eigenproblem

The methods described above generalize to the regular generalized eigenproblem $A-\lambda B$. We require regularity ($A-\lambda B$ square and $\det(A-\lambda B)$ not identically zero) since the eigenproblem is otherwise not well-posed. The concept of *invariant subspace* for the standard eigenproblem, a space $\mathbf{X}$ satisfying $A\mathbf{X}\subseteq\mathbf{X}$, is replaced by *a pair of deflating subspaces* $\mathbf{X}$ and $\mathbf{Y}$ of equal dimension such that $A\mathbf{X}+B\mathbf{X}=\mathbf{Y}$. Thus any algorithm is at least conceptually going to determine two spaces simultaneously. In this section we show how C, DMW and S generalize, how they are all equivalent to solving a generalized Riccati equation, and how the same convergence theory applies to all three as before.

First we show how to generalize Chatelin's algorithm C. We will seek two n by m matrices X and Y as bases for $\mathbf{X}$ and $\mathbf{Y}$. As before, we normalize X and Y by another full rank n by m matrix Z:

$$\begin{aligned} AX - Y(Z^*AX) &= 0 \qquad Z^*X = I \\ BX - Y(Z^*BX) &= 0 \quad , \quad Z^*Y = I \end{aligned} \tag{GenC}$$

As before we change bases (which means performing the same equivalence transformation on A and B) so that $Z = [I|0]^T$ and rewrite X and Y as

$$X = \begin{bmatrix} I \\ R \end{bmatrix} \quad , \quad Y = \begin{bmatrix} I \\ L \end{bmatrix} .$$

As before we call the transformed matrices A and B and partition them as

$$\begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} B_{11} & B_{21} \\ B_{12} & B_{22} \end{bmatrix}$$

with A_{11} and B_{11} being m by m and A_{22} and B_{22} being $n-m$ by $n-m$. If Z consists of orthonormal columns, this transformation may be taken to be orthonormal. Plugging into (GenC) yields

$$\begin{bmatrix} A_{22}R - LA_{11} = -A_{21} + LA_{12}R \\ B_{22}R - LB_{11} = -B_{21} + LB_{12}R \end{bmatrix} , \qquad \text{(GenC')}$$

the promised generalized Riccati equation.

There are at least two generalizations of DMW to the generalized eigenproblem. One might try to solve the equation

$$AX = BXC$$

where X is an n by m matrix whose columns span **X**, and C is an m by m matrix. The problem with this generalization of DMW is that it assumes BX is of full rank, since otherwise C might not exist. If the pencil $A-\lambda B$ has an infinite eigenvalue whose deflating subspace **X** we want, then BX will not be of full rank (numerically it will be nearly rank deficient) and consequently C will be very ill conditioned and hard to determine. If one knows B is well conditioned, however, this method could be used.

A better generalization of DMW is

$$AX = YA_Y \quad , \quad BX = YB_Y \qquad \text{(GenDMW)}$$

where X and Y are both n by m matrices and A_Y and B_Y are both m by m matrices. As in DMW, we will hold m rows each of X and Y constant, so that (GenDMW) consists of $2nm$ equations in $2nm$ unknowns.

Performing an equivalence transformation so that X and Y are of the form

$$X = \begin{bmatrix} I \\ R \end{bmatrix} \quad , \quad Y = \begin{bmatrix} I \\ L \end{bmatrix}$$

and plugging into (GenDMW) yields

$$\begin{matrix} A_{11} + A_{12}R = A_Y & & B_{11} + B_{12}R = B_Y \\ A_{21} + A_{22}R = LA_Y & , & B_{21} + B_{22}R = LB_Y \end{matrix} .$$

Substituting the expressions for A_Y and B_Y into the second equations yields

$$\begin{bmatrix} A_{22}R - LA_{11} = -A_{21} + LA_{12}R \\ B_{22}R - LB_{11} = -B_{21} + LB_{12}R \end{bmatrix} , \qquad \text{(GenDMW')}$$

the same generalized Riccati equation as before.

Stewart also deals with the generalized eigenproblem in [Stewart]. Again, this work was presented originally not as an algorithm but as a technique for doing perturbation theory. Stewart takes n by m matrices X and Y spanning approximate deflating subspaces as well as n by $n-m$ matrices X' and Y' spanning complementary subspaces so that $[X|X']$ and $[Y|Y']$ are nonsingular. The true deflating subspaces are represented as $X+X'R$ and $Y+Y'L$, where R and L are $n-m$ by m matrices to be determined. An equation for R and L is determined as follows. $X+X'R$ and $Y+Y'L$ will be deflating if and only if the lower left $n-m$ by m corners

of

$$[Y+Y'L|Y']^{-1} \cdot A \cdot [X+X'R|X'] = ([Y|Y'] \cdot \begin{bmatrix} I & 0 \\ L & I \end{bmatrix})^{-1} \cdot A \cdot ([X|X'] \cdot \begin{bmatrix} I & 0 \\ R & I \end{bmatrix})$$

and

$$[Y+Y'L|Y']^{-1} \cdot B \cdot [X+X'R|X'] = ([Y|Y'] \cdot \begin{bmatrix} I & 0 \\ L & I \end{bmatrix})^{-1} \cdot B \cdot ([X|X'] \cdot \begin{bmatrix} I & 0 \\ R & I \end{bmatrix})$$

are zero. As before, transform from the right and left so that $[X|X']$ and $[Y|Y']$ are identity matrices. if X and Y consist of orthonormal columns and X' and Y' span the orthogonal complements of the spaces spanned by X and Y, respectively, then this change of bases can be orthonormal. In this new bases we get that the lower left corners of

$$\begin{bmatrix} I & 0 \\ L & I \end{bmatrix}^{-1} \cdot \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \cdot \begin{bmatrix} I & 0 \\ R & I \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} I & 0 \\ L & I \end{bmatrix}^{-1} \cdot \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \cdot \begin{bmatrix} I & 0 \\ R & I \end{bmatrix}$$

must be zero, or

$$\begin{bmatrix} A_{22}R - LA_{11} = -A_{21} + LA_{12}R \\ B_{22}R - LB_{11} = -B_{21} + LB_{12}R \end{bmatrix} , \qquad \text{(GenS)}$$

the same generalized Riccati equation as before.

6. Solving the Generalized Riccati Equation

As for the earlier Riccati equation, there are two basic approaches to solving the generalized Riccati equation, the iteration

$$\begin{bmatrix} A_{22}R_{l+1} - L_{l+1}A_{11} = -A_{21} + L_lA_{12}R_l \\ B_{22}R_{l+1} - L_{l+1}B_{11} = -B_{21} + L_lB_{12}R_l \end{bmatrix} , \quad R_0 = 0 \ , \ L_0 = 0 \qquad \text{(GenIter)}$$

and Newton's method. Also as before, Stewart provides a convergence criterion for (GenIter) in terms of easily computable properties of the blocks A_{ij} and B_{ij}. A similar analysis leads to a slightly stronger criterion for the convergence of Newton's method.

To state Stewart's result we need to generalize the sep quantity used before. Following Stewart we define $\mathrm{dif}(A_{11},A_{22};B_{11},B_{22})$ as the smallest singular value of the linear operator that maps (R,L) to $(A_{22}R-LA_{11},B_{22}R-LB_{11})$. It is nonzero (i.e. the operator is invertible) as long as the pencils $A_{11}-\lambda B_{11}$ and $A_{22}-\lambda B_{22}$ have disjoint spectra. One can also show that it is the smallest singular value of the $2m(n-m)$ by $2m(n-m)$ matrix

$$\begin{bmatrix} I_m \otimes A_{22} & -A_{11}^T \otimes I_{n-m} \\ I_m \otimes B_{11} & -B_{11}^T \otimes I_{n-m} \end{bmatrix} .$$

Theorem 3: [Stewart] Let

$$\kappa = \frac{\|(A_{12},B_{12})\|_F \cdot \|(A_{21},B_{21})\|_F}{\mathrm{dif}^2(A_{11},A_{22};B_{11},B_{22})} .$$

Then if

$$\kappa < \frac{1}{4} \qquad \text{(GenLinearConvergence)}$$

(GenIter) converges linearly to a solution (R,L) which is the unique solution inside the ball

$$\|(R,L)\|_F \le \frac{1-(1-4\kappa)^{1/2}}{2\kappa} \cdot \frac{\|(A_{21},B_{21})\|_F}{\mathrm{dif}(A_{11},A_{22};B_{11},B_{22})} < 2 \cdot \frac{\|(A_{21},B_{21})\|_F}{\mathrm{dif}(A_{11},A_{22};B_{11},B_{22})} \ .$$

Furthermore, the convergence is linear with a contraction constant of no more than

$$1-(1-4\kappa)^{1/2} \ .$$

On comparing Theorems 1 and 3, we see they are almost identical, and indeed Stewart derives them both as special cases of a more general theorem which says when the iteration

$$Tx_{i+1} = g - \phi(x_i)$$

converges to a solution of $Tx = g - \phi(x)$, x a member of a Banach space, T an invertible linear operator, g a constant member of the Banach space, and ϕ a "quadratic" operator. Also, the interpretation of κ is similar: it is small if the initial approximate deflating subspaces are good approximations and if the spectrum associated with the desired deflating subspace is will separated from its complement.

Now we turn to the convergence of Newton's method. The same approach as before yields

$$\begin{aligned}(A_{22}-L_kA_{12})R_{k+1} - R_{k+1}(A_{11}+A_{12}R_k) &= -A_{21} - L_kA_{12}R_k \\ (B_{22}-L_kB_{12})R_{k+1} - R_{k+1}(B_{11}+B_{12}R_k) &= -B_{21} - L_kB_{12}R_k\end{aligned} \qquad \text{(GenNewton)}$$

a set of linear equations for (R_{k+1},L_{k+1}). The next theorem states a condition under which (GenNewton) converges:

Theorem 4: Let κ be defined as in Theorem 3. Then if

$$\kappa < \frac{1}{12} \qquad \text{(GenQuadraticConvergence)}$$

(GenNewton) will converge quadratically to the unique solution (R,L) inside the ball given in Theorem 3. If $E_k = R_k - R$ and $F_k = L_k - L$, then

$$\|(E_{k+1},F_{k+1})\|_F \le \frac{\|(A_{12},B_{12})\|_F}{\mathrm{dif}(A_{11},A_{22};B_{11},B_{22})} \cdot \frac{3}{2} \cdot \|(E_k,F_k)\|_F^2 \ .$$

Proof: The proof is analogous to the proof of Theorem 3. Take (GenNewton), subtract the same equation with R substituted for R_k and R_{k+1} and L substituted for L_k and L_{k+1}, and rearrange to obtain

$$\begin{aligned}(A_{22}-L_kA_{12})E_{k+1} - F_{k+1}(A_{11}+A_{12}R_k) &= -F_kA_{12}E_k \\ (B_{22}-L_kB_{12})E_{k+1} - F_{k+1}(B_{11}+B_{12}R_k) &= -F_kB_{12}E_k\end{aligned}$$

implying

$$\|(E_{k+1},F_{k+1})\|_F \le \frac{\|(E_k,F_k)\|_F^2 \cdot \|(A_{12},B_{12})\|_F}{\mathrm{dif}(A_{11}+A_{12}R_k,A_{22}-L_kA_{12};B_{11}+B_{12}R_k,B_{22}-L_kB_{12})} \ . \qquad \text{(Gen*)}$$

Thus, quadratic convergence is evident as long as the denominator is bounded away from 0. To prove this we need bounds on the growth of $\|(R_k,L_k)\|_F$.

From (GenNewton) we see that

$$\|(R_{k+1},L_{k+1})\|_F \le \frac{\|(A_{21},B_{21})\|_F + \|(A_{12},B_{12})\|_F \cdot \|(R_k,L_k)\|_F^2}{\mathrm{dif}(A_{11}+A_{12}R_k,A_{22}-L_kA_{12};B_{11}+B_{12}R_k,B_{22}-L_kB_{12})} \ .$$

From [Stewart] we have the lemma (GenLemma) that

$$\mathrm{dif}(A_{11}+E_{11},A_{22}+E_{22};B_{11}+F_{11},B_{22}+F_{22}) \geq \mathrm{dif}(A_{11},A_{22};B_{11},B_{22}) - 2\cdot\|(E_{11},E_{22},F_{11},F_{22})\|_F$$

which implies that

$$\|(R_{k+1},L_{k+1})\|_F \leq \frac{\|(A_{21},B_{21})\|_F + \|(A_{12},B_{12})\|_F\cdot\|(R_k,L_k)\|_F^2}{\mathrm{dif}(A_{11},A_{22};B_{11},B_{22}) - 2^{1/2}\cdot\|(A_{12},B_{12})\|_F\cdot\|(R_k,L_k)\|_F} .$$

Abbreviating

$$\|(R_{k+1},L_{k+1})\|_F = \frac{f_k\cdot\|(A_{21},B_{21})\|_F}{\mathrm{dif}(A_{11},A_{22};B_{11},B_{22})}$$

we see that

$$f_k \leq \frac{1+\kappa f_k^2}{1-2\kappa f_k}$$

with $f_0=0$. This is the same recurrence as in Theorem 3, and it is easy to see the rest of the proof follows as in the earlier theorem. Q.E.D.

7. The cost of Iteration versus Newton

In this section we discuss ways to solve the linear equations needed at each iteration of the algorithms, and show that (Iter) and (GenIter) cost only $O(n^2)$ arithmetic operations per iteration after a preprocessing cost of $O(n^3)$, whereas (Newton) and (GenNewton) cost $O(n^3)$ operations per iteration.

Let us consider the standard eigenproblem first. Iteration requires the solution of the set of linear equations

$$A_{22}R_{l+1} - R_{l+1}A_{11} = F(R_l)$$

for R_{l+1} at each iteration. This equation, called Sylvester's equation, has had solutions proposed in both [Bartels,Stewart] and [Golub,Nash,Van Loan]. Both algorithms have the property that after a preprocessing step (reducing A_{ll} to triangular or Hessenberg form) which takes $O(m^3+(n-m)^3)$ operations, it is possible to solve (Iter) for R_{l+1} in just $O((n-m)m^2+m(n-m)^2)$ operations. This savings is possible because the preprocessing reduces (Iter) to a logically triangular system. If $m\ll n$, a common case, preprocessing is $O(n^3)$ and solving is $O(n^2)$. Thus (Iter) is cheap after first preprocessing.

Newton, on the other hand, requires the solution of

$$(A_{22}-R_lA_{12})R_{l+1} - R_{l+1}(A_{11}+A_{12}R_l) = G(R_l)$$

at each step. This is also a Sylvester equation for R_{l+1}, but now the coefficient matrices $A_{22}-R_lA_{12}$ and $A_{11}+A_{12}R_l$ vary from step to step, making it necessary to preprocess at each step or at least frequently. Thus each step could take $O(n^3)$ operations instead of $O(n^2)$. This fact seems to make a modified Newton method in which the coefficient matrices are updated either only approximately or only occasionally attractive. In particular, it seems (Newton) could be solved iteratively for R_{l+1} using a quick solver of $A_{22}X-XA_{11}=Y$ as an approximate inverse for iterative refinement This would be especially true if the matrix were nearly block triangular so that R were small; then the operator $A_{22}X-XA_{11}$ would be a good approximation to $(A_{22}-R_lA_{12})X-X(A_{11}+A_{12}R_l)$. However, as we now show, we should not expect this scheme to be superior to using (Iter) for a while to get an approximate R, updating the matrix by transforming so that $[X+X'R|X']$ is the identity, and using (Iter) again. Letting S_i denote the ith iterate in the following iterative refinement algorithm for (Newton):

Algorithm 2:

1) Compute the residual $r_l = -A_{21}-R_lA_{12}R_l-(A_{22}-R_lA_{12})S_l+S_l(A_{11}+A_{12}R_l)$.

2) Solve $A_{22}d_l-d_lA_{11} = r_l$ for the correction d_l.

3) Update $S_{l+1}=S_l+d_l$.

After some manipulation we see

$$A_{22}S_{l+1}-S_{l+1}A_{11} = -A_{21}+R_lA_{12}R_l+R_lA_{12}(S_l-R_l)+(S_l-R_l)A_{12}R_l \quad \text{(ModNewton)}$$

Since we expect S_l to be a better approximation to R than R_l, we could let $R_l=S_l$ in the above formula. But then (ModNewton) would be identical to (Iter). This leads us to recommend the following version of the algorithm given earlier (as a matrix interpretation of Newton):

Algorithm 3:

1) Given bases X and X' for an approximate invariant subspace and a complementary space, transform the problem the problem so that $[X|X']=I$.

2) Take one or more steps of (Iter) with $R_0=0$, replace X by the better estimate $X+X'R$ and return to step 1).

The number of steps of (Iter) to take in step 2 above would depend on the convergence rate: if it is fast enough, there is no reason to return to step 1 and pay $O(n^3)$ operations.

The same considerations apply to the generalized eigenproblem. It turns out that if the A_{ll} and B_{ll} are triangular, the linear system

$$\begin{bmatrix} A_{22}R_{l+1} - L_{l+1}A_{11} = F_1(L_l,R_l) \\ B_{22}R_{l+1} - L_{l+1}B_{11} = F_2(L_l,R_l) \end{bmatrix}$$

is logically block triangular with 2 by 2 blocks, so that we can solve for the entries of L_{l+1} and R_{l+1} two at a time by substitution. (Analogs of both the [Bartels, Stewart] and [Golub,Nash,Van Loan]) algorithms are possible.) Thus again $O(n^3)$ operations of preprocessing allow each iteration of (GenIter) to cost only $O(n^2)$ iterations. Each iteration of (GenNewton), on the other hand, costs $O(n^3)$ as above, leading us to recommend an algorithm like Algorithm 3 above.

It is of interest to note the work on solving the Riccati equation in the control systems community ([Kleinman],[Arnold,Laub]). They are interested in solving variations of the Riccati equation (S), in particular when $A_{11}=-A_{22}^T$. Their standard approach is to transform to the corresponding matrix eigenvalue problem and use the QR algorithm. If more accuracy is needed, they use Newton. These algorithms have been implemented in a package of FORTRAN subroutine called RICPACK. Currently none of these algorithms apply to the general nonsymmetric version of the Riccati equation we consider in this paper, just the special form mentioned above.

8. Future Work

The approach taken in this paper does not tell us which of the algorithms presented has better numerical properties. In a future paper we intend to make numerical experiments comparing the speeds and accuracies of the algorithm of Dongarra, Moler and Wilkinson, the algorithm of Chatelin, and Algorithm 3 presented in the last section. In addition, the versions of these algorithms for the generalied eigenproblem will be programmed and compared.

9. References

[Arnold,Laub] W. Arnold, A. Laub, Generalized Eigenproblem Algorithms and Software for Algebraic Riccati Equations, Proc. IEEE, vol. 72, no. 12, Dec 1984

[Bartels,Stewart] R.H. Bartels, G. W. Stewart, Solution of the Matrix Equation $AX+XB=C$,

CACM, Vol. 15, No. 9, Sept. 1972, pp 820-826

[Chatelin] F. Chatelin, Simultaneous Newton's Iteration for the Eigenproblem, 1985

[Dongarra,Moler,Wilkinson] J. J. Dongarra, C. B. Moler, J. H. Wilkinson, Improving the Accuracy of Computed Eigenvalues and Eigenvectors, SIAM J. Num. Anal., Vol. 20, No. 1, Feb. 1983, pp 46-58

[Golub,Nash,Van Loan] G. Golub, S. Nash, C. Van Loan, A Hessenberg-Schur Method for the Problem $AX+XB=C$, IEEE Trans. Auto. Cntrl. Vol. AC-24, No. 6, Dec. 1979, pp 909-913

[Kleinman] D. L. Kleinman, On an iterative technique for Riccati equation computations, IEEE Trans. Auto. Contrl., vol. AC-13, pp 114-115, Feb 1968

[Stewart] G. W. Stewart, Error and Perturbation Bounds for Subspaces Associated with Certain Eigenvalue Problems, SIAM Review, Vol. 15, 1973, pp 752-764

C.1

NYU COMPSCI TR-185
Demmel, James Weldon
Three methods for refining
estimates of invariant
subspaces.

DATE DUE	BORROWER'S NAME